AI Synthesised Faces: Can Perceptual Training Improve Human Improve Detection?

Alexis McGuire[a,*], Sophie Nightingale[a] and Paul Taylor[a]

[a]Department of Psychology, Lancaster University, Lancaster LA1 4YW, United Kingdom
ORCID (Alexis McGuire): https://orcid.org/https://orcid.org/0009-0005-4481-2216, ORCID (Sophie Nightingale): https://orcid.org/https://orcid.org/0000-0002-6779-9203, ORCID (Paul Taylor): https://orcid.org/https://orcid.org/0000-0001-8743-7667

Abstract.

In this growing digital era, Artificial Intelligence (AI) is achieving new heights, and in parallel new threats arise. State-of-the-art models can now produce highly realistic synthetic faces that humans find hard to differentiate from real ones. For people who excel at this task, it remains unclear how they do this. We wondered if providing perceptual training, (in the form of tutorials designed to display visual pitfalls in AI-synthesised images) can improve human accuracy. Here, we investigate differences between two types of faces that are a product of AI: GAN-synthesised and diffusion-synthesised faces. We employ a human training approach, to investigate whether perceptual accuracy can be improved. We also examine which visual attributes participants associate with real and AI-synthesised faces. Example attributes assessed include familiarity and attractiveness, these have been linked to face processing theories such as face space theory. We investigated these attributes to understand if there are differences in processing mechanisms between real and AI-synthesised faces and if the faces differ in visual attributes. We also checked whether a relationship between participant confidence and classification accuracy exists. Findings suggest that synthetic images have reached a level of sophistication that consequently render simple human-centered training interventions ineffective.

1 Introduction

Synthetic faces are already highly realistic and when tasked with deciphering between real and AI-synthesised faces, human performance is poor (at around chance level) [1, 12, 11]. These limits in human detection abilities emerge as a challenge, particularly when synthesised faces are used maliciously; for example, to facilitate romance fraud [2]. AI synthesised faces can be utilised by scammers to create fake social media profiles, build relationships with victims with the aim of exploiting them for financial benefit. Currently, limits in perceptual ability to tell real images apart from fake ones make humans particularly vulnerable to this type of deception. In addition, the low barriers to create these faces means that almost anyone, even those without technical skills, can now create a fake image. This is especially worrying for newer diffusion models [3], whose output can be guided using simple linguistic text prompts, in addition to image prompts.

* Corresponding Author. Email: l.mcguire@lancaster.ac.uk

Considering the points raised above, there is a need to understand whether human perceptual abilities can be enhanced—that is, are there ways to improve human ability to identify real and fake images? Many computational approaches have been trialed, one method is fingerprinting, this can serve to distinguish AI-synthesised images from real ones [14]. Fingerprinting is a technique that involves embedding patterns in the image at the creation stage, these can be detected by computers to show provenance information, signifying if the image is likely real or AI-synthesised. These fingerprints are not apparent to the human eye, hence visually the image remains the same. Without external influence leading to doubts about authenticity of the image, it is unlikely the viewer will engage in additional measures to reveal this hidden information. Furthermore, it is unlikely that most ordinary people will be able to access these watermarks, hence unless suspicions surrounding authenticity are raised to a wider level, this approach is not universal, or accessible for the average person. Therefore, for the everyday viewer it is important to understand what makes these faces so realistic and if a method exists to increase perceptual performance.

Media literacy campaigns can help to inoculate the public from a human visual perspective [7], these are a variety of taught strategies that people can use to help inform their decisions surrounding online engagement, for example, debunking strategies [6]. However, measuring the impact of these interventions can be difficult as exposure to content often does not guarantee active engagement; for example, unless the program is particularly engaging and tailored to a specific audience, a change in behaviour is unlikely. Specific human focused training techniques on how to identify AI-synthesised images are therefore a favourable alternative, to increase engagement, and this is a measurable intervention. There has been limited research surrounding human training interventions with the aim of increasing synthetic face detection. One study found that for GAN-synthesised faces, perceptual training to help humans identify rendering artifacts in conjunction with trial-by-trial feedback led to an increase in accuracy [11], however, despite this increase, average accuracy was still poor. Another study highlighted a slight increase in accuracy for participants who undertook training, indicating common artifacts in GAN-synthesised faces [10]. However, it is not yet clear what impact perceptual training has for diffusion-synthesised images and whether these images differ visually to the already highly realistic precursor, GAN-synthesised. For example, it is unclear what specifically within these synthetic faces makes them so human-like; are there visual at-

tributes that stand out for AI-synthesised faces? Previous research aimed to answer this question for GAN-synthesised faces; Miller [9] discovered that faces rated by participants as more familiar were more likely to be judged as human. This result aligns with theories of face processing: face space theory [13]. This theory attunes to how humans store facial information in a multidimensional space. In this space, faces more typical to the facial norm are clustered together, and more distinctive faces are placed further apart. Miller et al [9] evidenced that faces rated lower for memorability and attractiveness had a greater chance of being labeled as human than AI-generated. Compared to real faces, GAN-synthesised faces were rated as more average, familiar, and attractive. The results suggest that humans are observing subtle perceptual differences between real and GAN-synthesised faces. We aimed to replicate this finding and test whether the attributes that people identify vary between GAN and diffusion-synthesised. This examination will help to identify which attributes are related to higher levels of realism for AI-generated faces.

The high realism of these AI-generated images mixed with a human predisposition to demonstrate overconfidence in their decisions [5] can be costly. Research has shown that people are overconfident in their ability to detect deepfake videos [4] and also GAN-synthesised faces [9]. This overconfidence may make individuals more prone to deception. This current research employs a human training intervention which aims to educate participants on the common artifacts in AI faces (that typically result from the synthesis process) and to investigate whether this can be used to increase detection performance. We aim to examine whether there are differences across two types of AI-synthesised faces in terms of visual attributes. Considering specific facial attributes will help to understand any differences in face perception for real and AI-generated faces, for example, what attributes do participants rely on to make judgments of real vs AI-synthesised, and in turn, can we use this information to guide future interventions. This research also aims to investigate participant confidence accuracy relationships, to inform whether humans have insights into their ability to detect AI-synthesised faces.

2 Research Objectives and Questions

In previous work, we have evidenced that AI-synthesised faces are highly realistic and trustworthy, even more so than real faces [8]. The research objectives for this study are as follows: 1) to investigate the impact of perceptual training on human ability to distinguish AI-synthesised images from real ones, 2) to distinguish what attributes people associate with real and AI (GAN and diffusion) faces, and 3) to examine the relationship between participant confidence and accuracy when distinguishing between real and AI-generated faces.

My thesis aims to answer the following research questions: 1.) Can perceptual training interventions improve the ability of humans to detect AI-generated faces. 2.) How do humans perceive AI faces, does this differ across faces produced by differing AI models. 3.) Are AI faces suitable for use within police investigation, specifically for eyewitness testimony.

3 Preliminary Results

The full data is yet to be analysed; initial analysis indicates that overall participants classified all images at just above chance performance (57.6%, 95% CI [56.5%, 58.7%]), (d'= 0.29, 95% CI [0.23,0.36]), with a bias to respond AI-generated (c= -0.22, 95% CI [-0.27, -0.18]. These results indicate that participants could not reliably distinguish between the real and AI faces. With respect to the training intervention, average accuracy was significantly higher in the training group (62.0%, 95% CI [60.1%, 64.0%]) compared to the control group (55.0%, 95% CI [53.1%, 56.0%]), t(309.8)= -5.59, p=<0.01). Despite a significant increase in accuracy for the training group, we found no significant change in discriminability. In the training condition, average d' = 0.34, 95% CI [0.22, 0.46] whereas in the control condition discriminability was lower d' = 0.26 95% CI [0.16, 0.37] — this difference is not statistically significant t(298.27) = -0.91, p = 0.36). There is, however, a significant criterion shift: participants who received training had a significantly lower criterion value c= -0.56 95% CI [-0.64, -0.48] compared to the control group c= 0.01 95% CI[-0.06, 0.08] this difference was significant t(293.18)= 10.44, p<0.01). This suggests that participants who received training were more likely to respond AI-synthesised than those in the control condition. Additional planned analyses include a mixed-effects logistic regression to model classification accuracy, with training, face type and attributes as fixed effects, also including random effects. We will also employ a lens model to investigate the relationships between face types and attribute ratings, and also a confidence- accuracy analysis.

4 Future Work

Future work will investigate the impact of participant race on accuracy and the application of AI-synthesised faces for use within police investigation.

Acknowledgements

This PhD is supervised by Dr. Sophie Nightingale and Prof. Paul Taylor. We would like to thank the Centre for Research and Evidence on Security Threats for providing funding for this research. (ESRC Award: ES/V002775/1), which is funded in part by the UK Home Office and security and intelligence agencies (see the public grant decision here: https://gtr.ukri.org/projects?ref=ES%2FV002775%2F1). The funding arrangements required this paper to be reviewed to ensure that its contents did not violate the Official Secrets Act nor disclose sensitive, classified and/or personal information.

References

[1] S. D. Bray, S. D. Johnson, and B. Kleinberg. Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity*, 9 (1):tyad011, 2023.

[2] C. Cross. Using artificial intelligence (ai) and deepfakes to deceive victims: the need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety*, 24(1):30–41, 2022.

[3] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[4] N. C. Köbis, B. Doležalová, and I. Soraperra. Fooled twice: People cannot detect deepfakes but think they can. *Iscience*, 24(11), 2021.

[5] J. Kruger and D. Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[6] S. Lewandowsky, J. Cook, U. Ecker, D. Albarracin, M. A. Amazeen, P. Kendou, D. Lombardi, E. Newman, G. Pennycook, E. Porter, et al. *The debunking handbook 2020*. 2020.

[7] A. M. McGowan-Kirsch and G. V. Quinlivan. Educating emerging citizens: Media literacy as a tool for combating the spread of image-based misinformation. *Communication Teacher*, 38(1):41–52, 2024.

[8] A. McGuire, M. Bohacek, H. Farid, P. Taylor, and S. Nightingale. How realistic are ai-generated faces? Aug. 2024. Perception, ECVP 2024 abstracts PDF, p. 202.

[9] E. J. Miller, B. A. Steward, Z. Witkower, C. A. Sutherland, E. G. Krumhuber, and A. Dawel. Ai hyperrealism: Why ai faces are perceived as more real than human ones. *Psychological science*, 34(12): 1390–1403, 2023.

[10] N. B. Mohamed, G. Bogdanel, and H. G. Moreno. Is training useful to detect deepfakes?: A preliminary study. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5. IEEE, 2023.

[11] S. J. Nightingale and H. Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

[12] B. Shen, B. RichardWebster, A. O'Toole, K. Bowyer, and W. J. Scheirer. A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[13] T. Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, 43(2):161–204, 1991.

[14] N. Yu, L. S. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.